What are the implications of Curriculum Learning strategy on IRL methods?

Mikhail Vlasenko, Luciano Cavalcante Siebert, Angelo Caregnato Neto

EEMCS, TU Delft

Introduction

Inverse Reinforcement Learning (IRL) is an aspect of machine learning that enables artificial agents to infer the reward function from observed expert demonstrations. Adversarial IRL (AIRL) is a promising algorithm that is postulated to recover non-linear rewards in environments with unknown dynamics.
Curriculum Learning (CL) is a learning strategy, inspired by the human learning process. It has been shown to accelerate convergence and improve generalization in Reinforcement Learning (RL). This study investigates the potential benefits of applying the CL strategy to the AIRL algorithm.

Background

Fu et al. [1] introduced an adversarial training approach to IRL in their paper about AIRL. The authors show that rewards recovered by AIRL generalize better than those produced by previous methods and, crucially, are more robust to changes in the environment during training. Unlike previous methods, AIRL has been shown to perform well even in high-dimensional control tasks.
Through an iterative process, AIRL strives to train a good policy for the environment. The incorporation of the reward model in the adversarial setup allows for the learning of sophisticated and resilient reward functions. Figure 1 schematically presents the information flow in the algorithm.

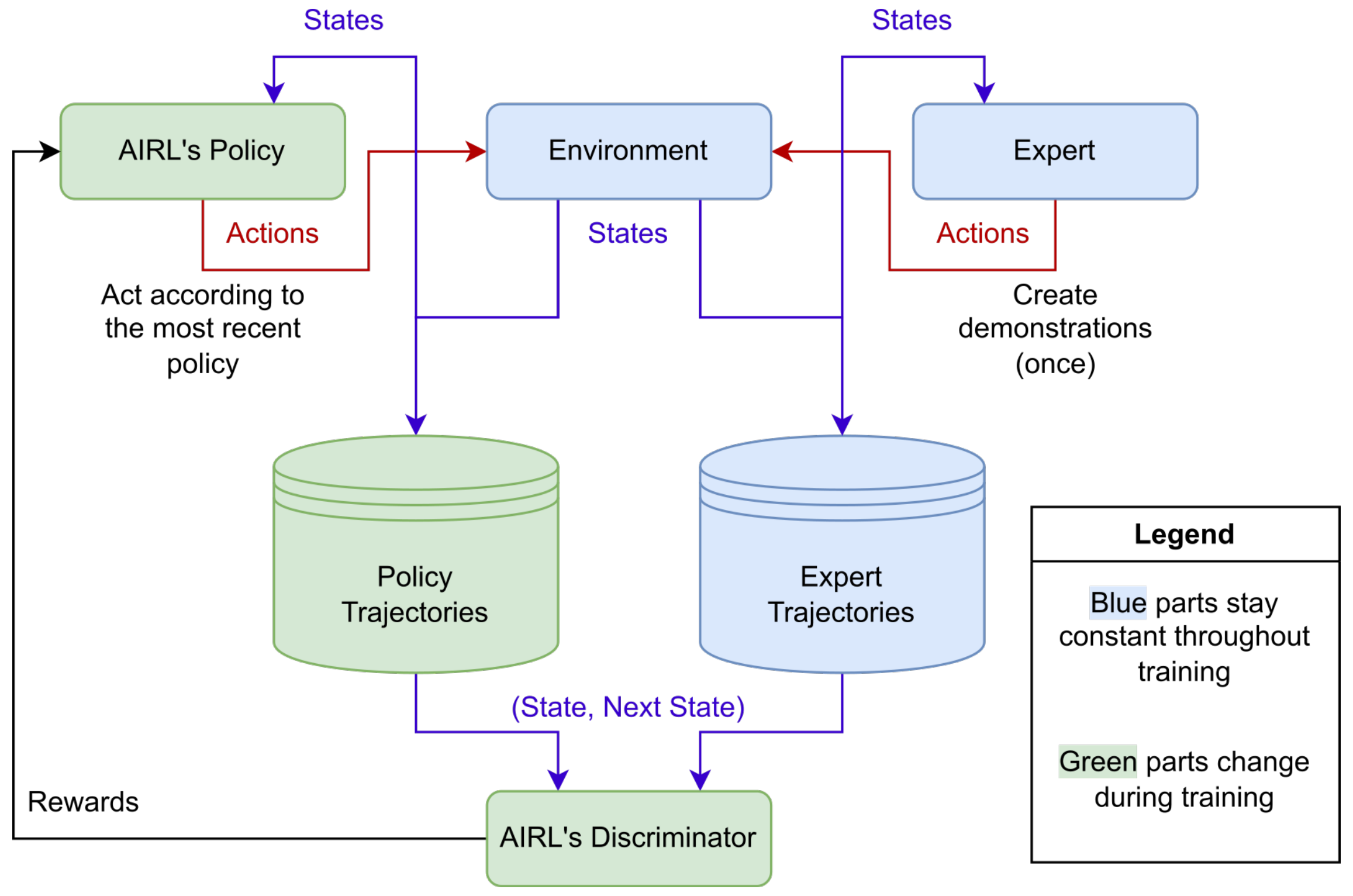

Figure 1:AIRL algorithm

Method

To apply CL to AIRL, we first construct a curriculum - a set of environments that will be used for training the policy and discriminator, and how many steps they are used for.

Algorithm 1 CL for AIRL

```
procedure TRAIN(environments, steps)
    $\pi$, $\mathcal{D}$ ← Randomly Initialize AIRL
    for env in environments do
        for i in 0..steps[env] do
            $\pi$, $\mathcal{D}$ ← train(env, $\pi$, $\mathcal{D}$)
        end for
    end for
    return $\pi$, $\mathcal{D}$
end procedure
```

For our experiments, we use a randomized partially observable Markov decision process in the form of a grid-world-like environment. The goal of the agent is to reach the target in the given amount of turns. Additionally, the agent receives a reward between -1 and 1 for stepping on an unvisited tile.

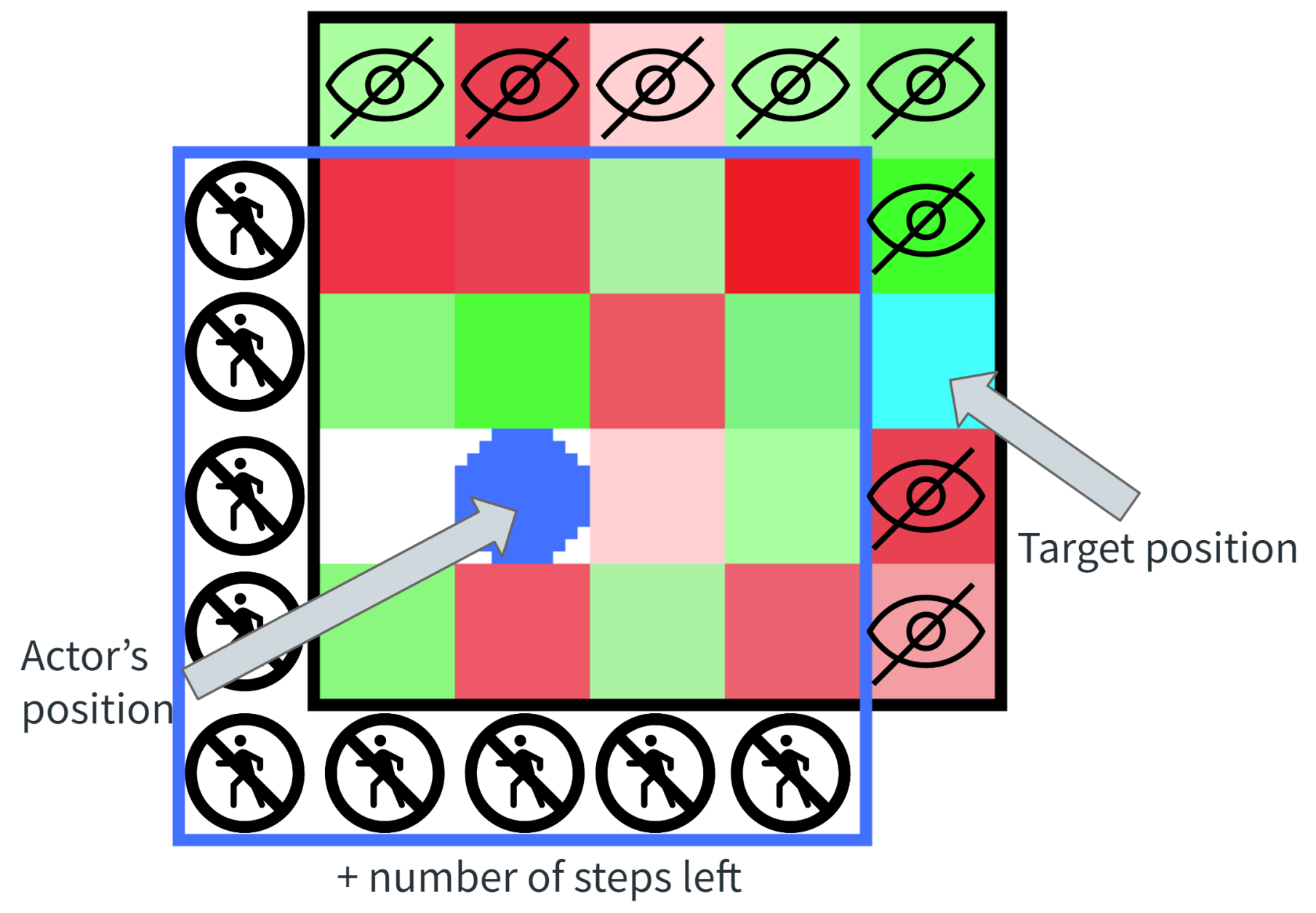

Figure 2:Visualization of observation space in the environment

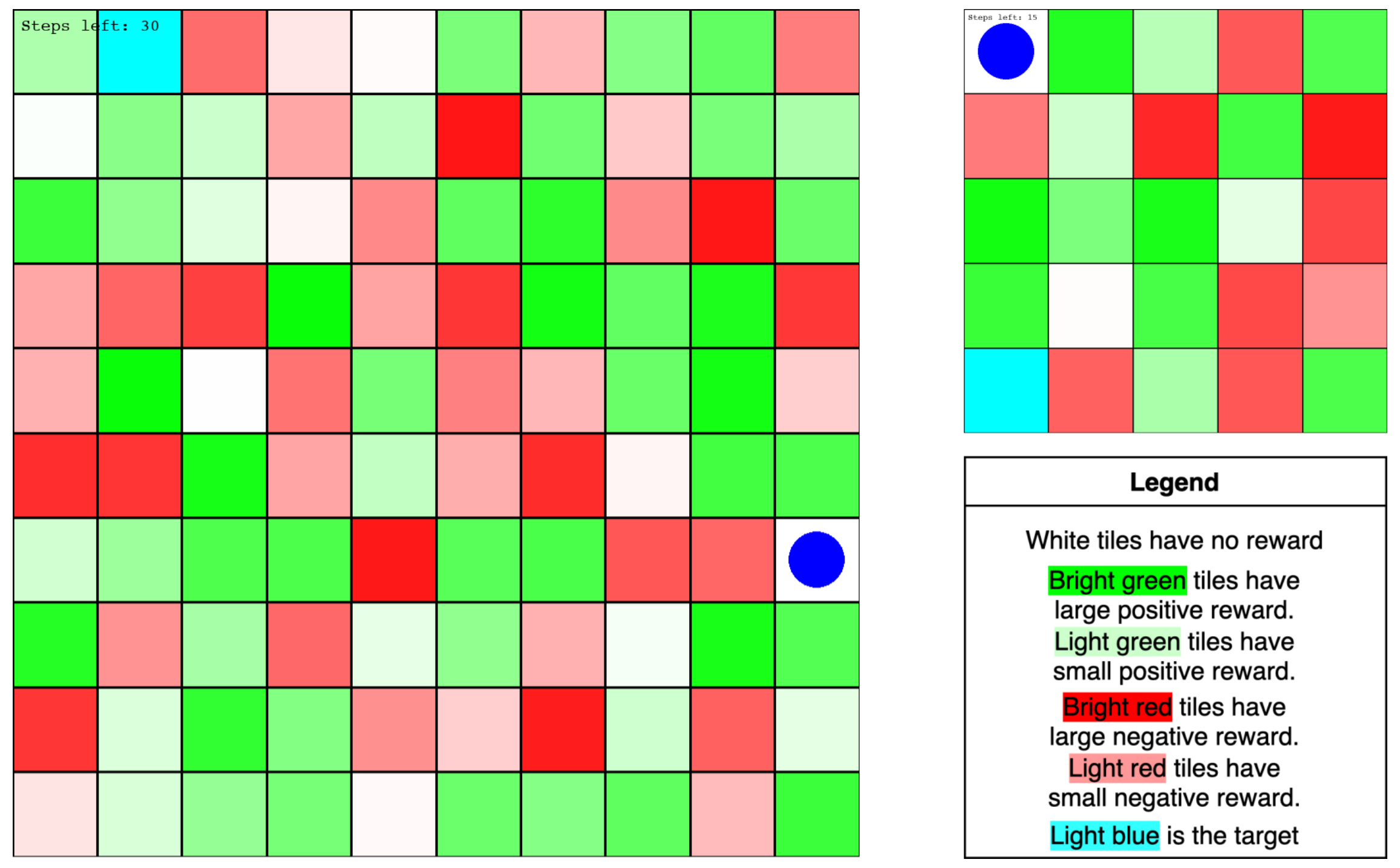
Steps left: 30
Steps left: 15
Legend
White tiles have no reward
Bright green tiles have large positive reward.
Light green tiles have small positive reward.
Bright red tiles have large negative reward.
Light red tiles have small negative reward.
Light blue is the target

Figure 3:Example starting states in the environment

Results

For evaluation, we compute the following metric from the latest AIRL policy trajectories in the evaluation environment.

$$\bar{R}_{true} = \frac{1}{N_{episodes}} \sum_{i=1}^{N_{episodes}} \sum_{t=0}^{T_{eval}} r_{i,t}$$

The *increasing grid size* curriculum was found to work best for the designed environment. For our goal configuration, we use $size = 10$. In the curriculum, the model starts training in an environment with $size = 5$, and only then is transferred the $size = 10$ configuration.

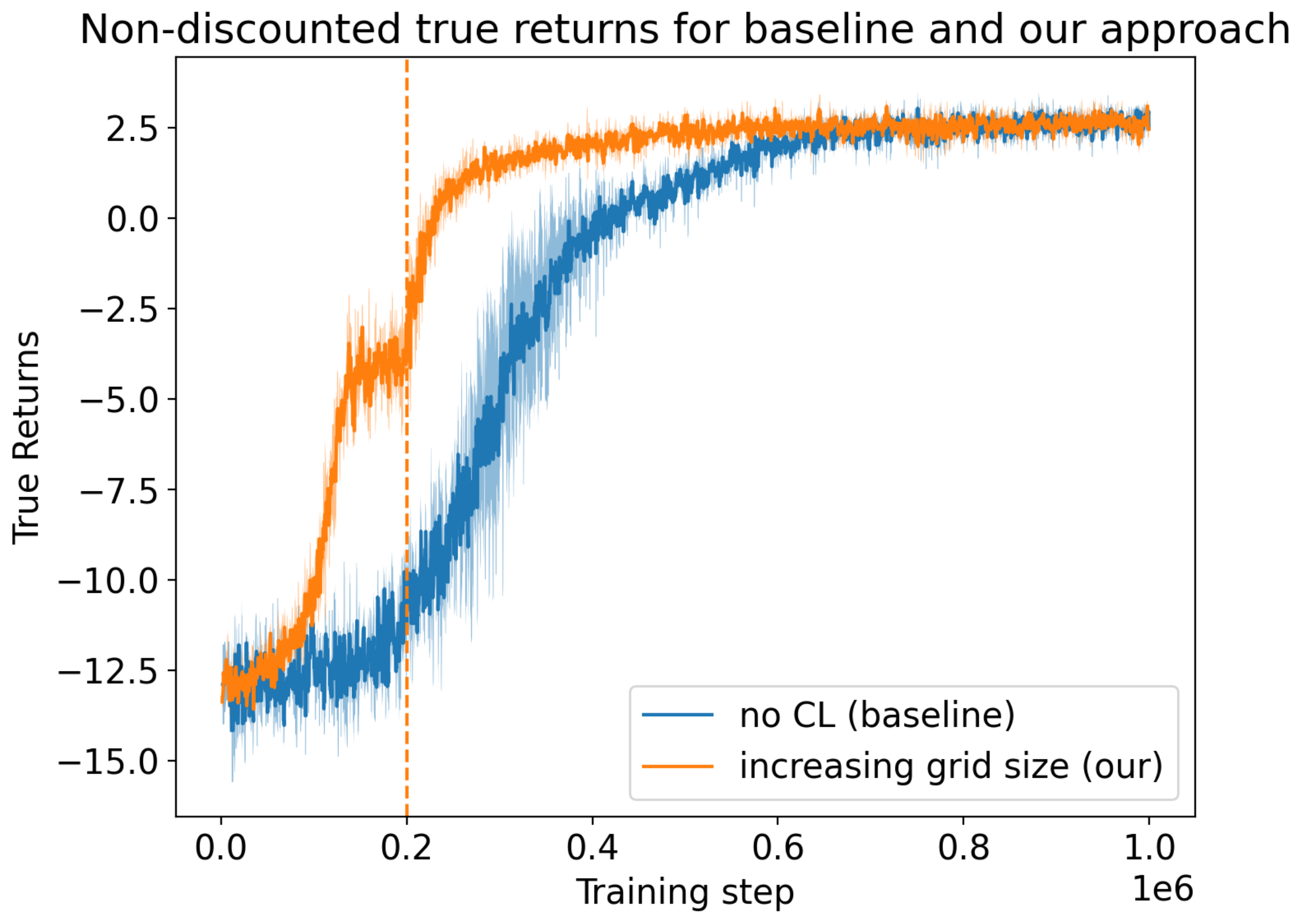

Figure 4:Rewards of runs with and without curriculum learning

We also experiment with moving from the $size = 5$ configuration at different time steps.

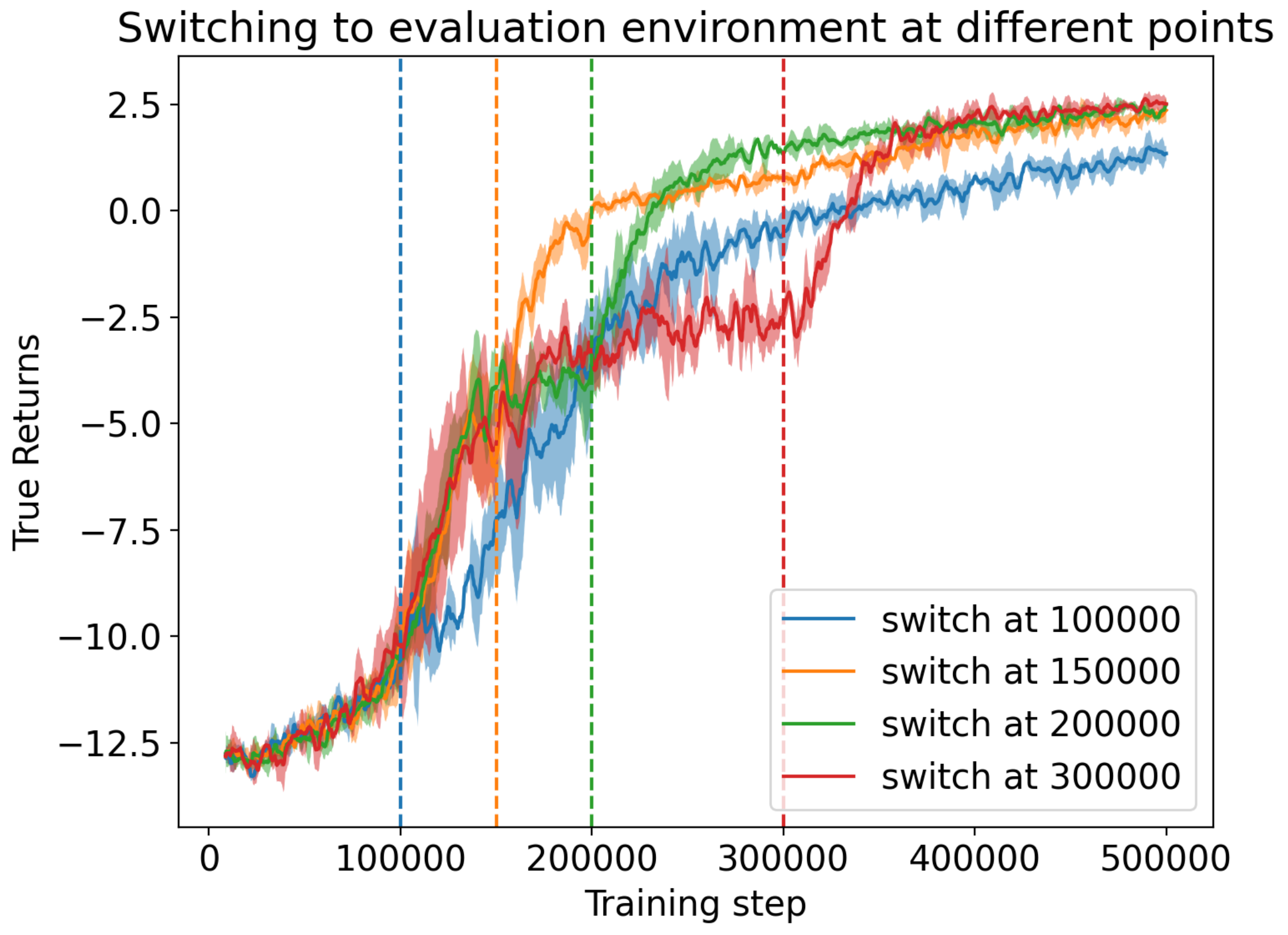

Figure 5:Different environment switch points

Results

CL can also be useful if the amount of expert demonstrations is scarce. Pretraining in a different environment, where expert demonstrations are easier to gather, proves beneficial for AIRL.
For the following experiment, we reduce the amount of expert data to 50 environment steps. When the model trains with CL, it starts with 50 expert data points in the environment with $size = 5$, and only then switches to the evaluation setting.

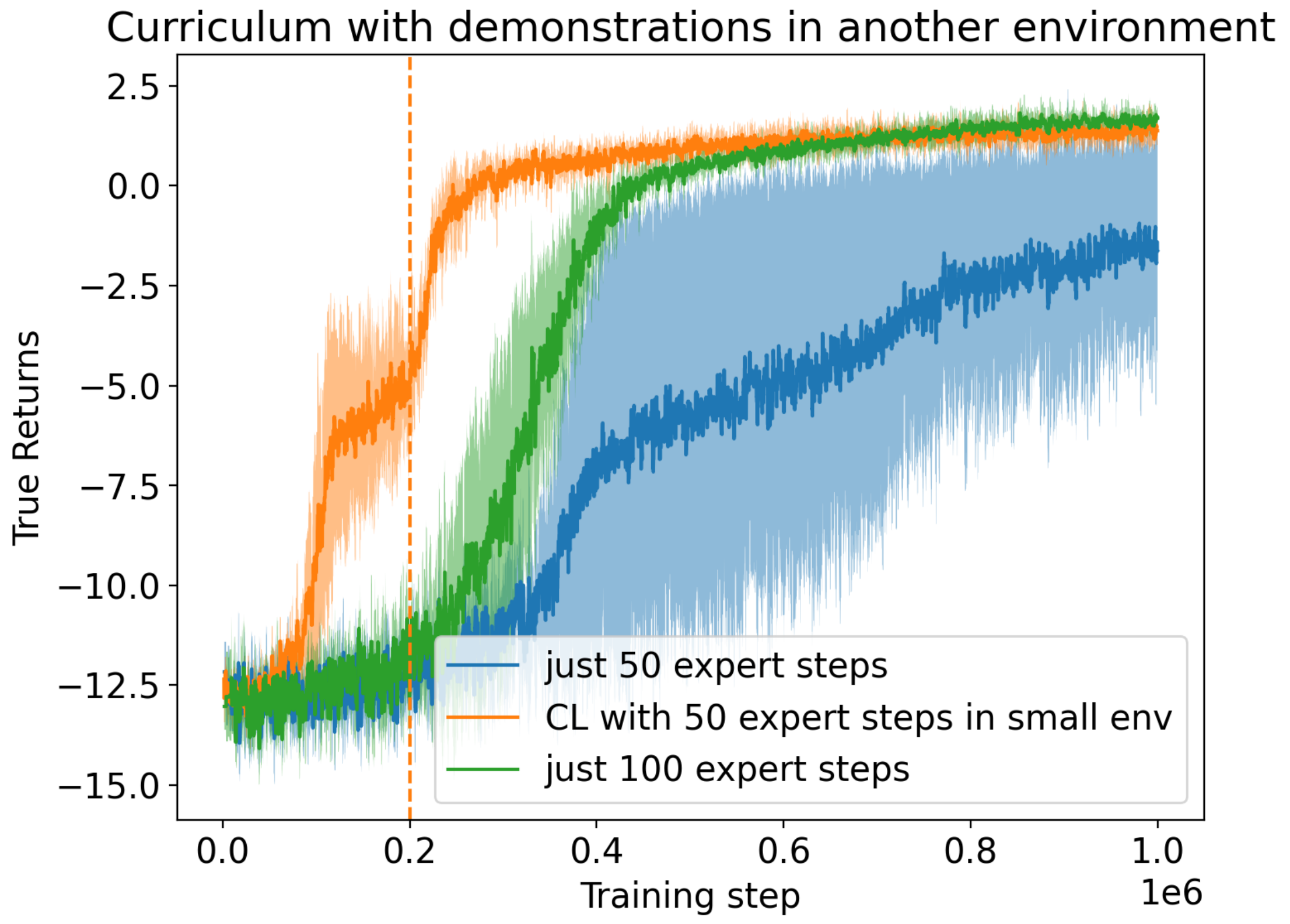

Figure 6:Curriculum with demos in another environment

Conclusion

Our results show, that a well-constructed curriculum can enhance the performance of AIRL twofold in both key aspects: the speed of convergence and the efficiency of using expert demonstrations. We thus conclude that CL can be a useful addition to an AIRL-based solution.

References

[1] Justin Fu, Katie Luo, and Sergey Levine.
Learning robust rewards with adversarial inverse reinforcement learning.

[2] Andrew Y Ng, Stuart Russell, et al.
Algorithms for inverse reinforcement learning.

[3] Yoshua Bengio et al.
Curriculum learning.